

## The Techniques of Mass Collaboration: A Third Way Out

I'm not the first [to suggest that the Internet could be used for bringing users together to build grand databases](#). The most famous example is the Semantic Web project (where, in full disclosure, I worked for several years). The project, spearheaded by Tim Berners-Lee, inventor of the Web, proposed to extend the working model of the Web to more structured data, so that instead of simply publishing text web pages, users could publish their own databases, which could be aggregated by search engines like Google into major resources.

The Semantic Web project has received an enormous amount of criticism, much (in my view) rooted in misunderstandings, but much legitimate as well. In the news today is just the most recent example, in which famed computer scientist turned Google executive [Peter Norvig challenged Tim Berners-Lee](#) on the subject at a conference.

The confrontation symbolizes the (at least imagined) standard debate on the subject, which Mark Pilgrim termed [million dollar markup versus million dollar code](#). Berners-Lee's W3C, the supposed proponent of million dollar markup, argues that users should publish documents that state in special languages that computers can process exactly what they want to say. Meanwhile Google, the supposed proponent of million dollar code, thinks this is an impractical fantasy, and that the only way forward is to write more advanced software to try to extract the meaning from the messes that users will inevitably create.<sup>[1]</sup>

[1]: I say supposed because although this is typically how the debate is seen, I don't think either the W3C or Google actually hold the strict positions on the subject typically ascribed to them. Nonetheless, the question is real and it's convenient to consider the strongest forms of the positions.

But yesterday I suggested what might be thought of as a third way out; one Pilgrim might call million dollar users. Both the code and the markup positions make the assumption that users will be publishing their own work on their own websites and thus we'll need some way of reconciling it. But Wikipedia points to a different model, where all the users come to *one* website, where the interface for inputting data in the proper format is clear and unambiguous, and the users can work together to resolve any conflicts that may come up.

Indeed, this method strikes me as so superior that I'm surprised I don't see it discussed in this context more often. Ignorance doesn't seem plausible; even if Wikipedia was a late-comer, sites like [ChefMoz](#) and [MusicBrainz](#) followed this model and were Semantic Web case studies. (Full disclosure: I worked on the Semantic Web portions of MusicBrainz.) Perhaps the reason is simply that both sides — W3C and Google — have the existing Web as the foundation for their work, so it's not surprising that they assume future work will follow from the same basic model.

One possible criticism of the million dollar users proposal is that it's somehow less free than the individualist approach. One site will end up being in charge of all the data and thus will be able to control its formation. This is perhaps not ideal, certainly, but if the data is made available under a free license it's no worse than things are now with free software. Those angry with the policies can always exercise their right to "fork" the project if they don't like the direction things are going. Not ideal, certainly, but we can try to dampen such problems by making sure the central sites are run as democratically as possible.

Another argument is that innovation will be hampered: under the individualist model, any person can start doing a new thing with their data, and hope that others will pick up the technique. In the centralized model, users are limited by the functionality of the centralized site. This too can be ameliorated by making the centralized site as open to innovation as possible, but even if it's closed, other people can still do new things by downloading the data and building additional services on top of it (as indeed [many have done with Wikipedia](#)).

It's been eight years since Tim Berners-Lee published his [Semantic Web Roadmap](#) and it's difficult to deny that things aren't exactly going as planned. Actual adoption of Semantic Web technologies has been negligible and nothing that promises to change that appears on the horizon. Meanwhile, the million dollar code people have not fared much better. Google has been able to launch a handful of very targeted features, like [music search](#) and [answers to very specific kinds of questions](#) but these are mere conveniences, far from changing the way we use the Web.

By contrast, Wikipedia has seen explosive growth, Amazon.com has become the premier site for product information, and when people these days talk about user-generated content, they don't even consider the individualized sense that the W3C and Google assume. Perhaps it's time to try the third way out.

*You should follow me on twitter [here](#).*

July 19, 2006